Best Practices in Quantitative Methods

Edited by Jason W. Osborne

SAGE

3

Best Practices in Interrater Reliability

Three Common Approaches

Steven E. Stemler

Jessica Tsai

The concept of interrater reliability[1] permeates many facets of modern society. For example, court cases based on a trial by jury require unanimous agreement from jurors regarding the verdict, life-threatening medical diagnoses often require a second or third opinion from health care professionals, student essays written in the context of high-stakes standardized testing receive points based on the judgment of multiple readers, and Olympic competitions, such as figure skating, award medals to participants based on quantitative ratings of performance provided by an international panel of judges.

Any time multiple judges are used to determine important outcomes, certain technical and procedural questions emerge. Some of the more common questions are as follows: How many raters do we need to be confident in our results? What is the minimum level of agreement that my raters should achieve? And is it necessary for raters to agree exactly, or is it acceptable for them to differ from each other so long as their difference is systematic and can therefore be corrected?

Key Questions to Ask Before Conducting an Interrater Reliability Study

If you are at the point in your research where you are considering conducting an interrater reliability study, then there are three important questions worth considering:

1. What is the purpose of conducting your interrater reliability study?
2. What is the nature of your data?
3. What resources do you have at your disposal (e.g., technical expertise, time, money)?

The answers to these questions will help determine the best statistical approach to use for your study.

What Is the Purpose of Conducting Your Interrater Reliability Study?

There are three main reasons why people may wish to conduct an interrater reliability study. Perhaps the most popular reason is that the researcher is interested in getting a single final score on a variable (such as an essay grade) for use in subsequent data analysis and statistical modeling but first must prove that the scoring is not "subjective" or "biased." For example, this is often the goal in the context of educational testing where large-scale state testing programs might use multiple raters to grade student essays for the ultimate purpose of providing an overall appraisal of each student's current level of academic achievement. In such cases, the documentation of interrater reliability is usually just a means to an end—the end of creating a single summary score for use in subsequent data analyses—and the researcher may have little inherent interest in the details of the interrater reliability analysis per se. This is a perfectly acceptable reason for wanting to conduct an interrater reliability study; however, researchers must be particularly cautious about the assumptions they are making when summarizing the data from multiple raters to generate a single summary score for each student. For example, simply taking the mean of the ratings of two independent raters may, in some circumstances, actually lead to biased estimates of student ability, even when the scoring by independent raters is highly correlated (we return to this point later in the chapter).

A second common reason for conducting an interrater reliability study is to evaluate a newly developed scoring rubric to see if it is "working" or if it needs to be modified. For example, one may wish to evaluate the accuracy of multiple ratings in the absence of a "gold standard." Consider a situation in which independent judges must rate the creativity of a piece of artwork. Because there is no objective rule to indicate the "true" creativity of a piece of art, a minimum first step in establishing that there is such a thing as creativity is to demonstrate that independent raters can at least reliably classify objects according to how well they meet the assumptions of the construct. Thus, independent observers must subjectively interpret the work of art and rate the degree to which an underlying construct (e.g., creativity) is present. In situations such as these, the establishment of interrater reliability becomes a goal in and of itself. If a researcher is able to demonstrate that independent parties can reliably rate objects along the continuum of the construct, this provides some good objective evidence for the existence of the construct. A natural subsequent step is to analyze individual scores according to the criteria.

Finally, a third reason for conducting an interrater reliability study is to validate how well ratings reflect a known "true" state of affairs (e.g., a validation study). For example, suppose that a researcher believes that he or she has developed a new colon cancer screening technique that should be highly predictive. The first thing the researcher might do is train another provider to use the technique and compare the extent to which the independent rater agrees with him or her on the classification of people who have cancer and those who do not. Next, the researcher might attempt to predict the prevalence of cancer using a formal diagnosis via more traditional methods (e.g., biopsy) to compare the extent to which the new technique is accurately predicting the diagnosis generated by the known technique. In other words, the reason for conducting an interrater reliability study in this circumstance is because it is not enough that independent raters have high levels of interrater reliability; what really matters is the level of reliability in predicting the actual occurrence of cancer as compared with a "gold standard"—in this case, the rate of classification based on an established technique.

Once you have determined the primary purpose for conducting an interrater reliability study, the next step is to consider the nature of the data that you have or will collect.

What Is the Nature of Your Data?

There are four important points to consider with regard to the nature of your data. First, it is important to know whether your data are considered nominal, ordinal, interval, or ratio (Stevens, 1946). Certain statistical techniques are better suited to certain types of data. For example, if the data you are evaluating are nominal (i.e., the differences between the categories you are rating are qualitative), then there are relatively few statistical methods for you to choose from (e.g., percent agreement, Cohen's kappa). If, on the other hand, the data are measured at

the ratio level, then the data meet the criteria for use by most of the techniques discussed in this chapter.

Once you have determined the type of data used for the rating scale, you should then examine the distribution of your data using a histogram or bar chart. Are the ratings of each rater normally distributed, uniformly distributed, or skewed? If the rating data exhibit restricted variability, this can severely affect consistency estimates as well as consensus-based estimates, threatening the validity of the interpretations made from the interrater reliability estimates. Thus, it is important to have some idea of the distribution of ratings in order to select the best statistical technique for analyzing the data.

The third important thing to investigate is whether the judges who rated the data agreed on the underlying trait definition. For example, if two raters are judging the creativity of a piece of artwork, one rater may believe that creativity is 50% novelty and 50% task appropriateness. By contrast, another rater may judge creativity to consist of 50% novelty, 35% task appropriateness, and 15% elaboration. These differences in perception will introduce extraneous error into the ratings. The extent to which your raters are defining the construct in a similar way can be empirically evaluated using measurement approaches to interrater reliability (e.g., factor analysis, a procedure that is further described later in this chapter).

Finally, even if the raters agree as to the structure, do they assign people into the same category along the continuum, or does one judge assign a person "poor" in mathematics while another judge classifies that same person as "good"? In other words, are they using the rating categories the same way? This can be evaluated using consensus estimates (e.g., via tests of marginal homogeneity).

After specifying the purpose of the study and thinking about the nature of the data that will be used in the analysis, the final question to ask is the pragmatic question of what resources you have at your disposal.

What Resources Do You Have at Your Disposal?

As most people know from their life experience, "best" does not always mean most expensive or most resource intensive. Similarly, within the context of interrater reliability, it is not always necessary to choose a technique that yields the maximum amount of information or that requires sophisticated statistical analyses in order to gain useful information. There are times when a crude estimate may yield sufficient information—for example, within the context of a low-stakes, exploratory research study. There are other times when the estimates must be as precise as possible—for example, within the context of situations that have direct, important stakes for the participants in the study.

The question of resources often has an influence on the way that interrater reliability studies are conducted. For example, if you are a newcomer who is running a pilot study to determine whether to continue on a particular line of research, and time and money are limited, then a simpler technique such as the percent agreement, kappa, or even correlational estimates may be the best match. On the other hand, if you are in a situation where you have a high-stakes test that needs to be graded relatively quickly, and money is not a major issue, then a more advanced measurement approach (e.g., the many-facets Rasch model) is most likely the best selection.

As an additional example, if the goal of your study is to understand the underlying nature of a construct that to date has no objective, agreed-on definition (e.g., wisdom), then achieving consensus among raters in applying a scoring criterion will be of paramount importance. By contrast, if the goal of the study is to generate summary scores for individuals that will be used in later analyses, and it is not critical that raters come to exact agreement on how to use a rating scale, then consistency or measurement estimates of interrater reliability will be sufficient.

Summary

Once you have answered the three main questions discussed in this section, you will be in a much better position to choose a suitable technique for your project. In the next section of this chapter, we will discuss (a) the most popular statistics used to compute interrater reliability, (b) the computation and interpretation of the results of statistics using worked examples, (c) the implications for summarizing data that follow from each technique, and (d) the advantages and disadvantages of each technique.

Choosing the Best Approach for the Job

Many textbooks in the field of educational and psychological measurement and statistics (e.g., Anastasi & Urbina, 1997; Cohen, Cohen, West, & Aiken, 2003; Crocker & Algina, 1986; Hopkins, 1998; von Eye & Mun, 2004) describe interrater reliability as if it were a unitary concept lending itself to a single, "best" approach across all situations. Yet, the methodological literature related to interrater reliability constitutes a hodgepodge of statistical techniques, each of which provides a particular kind of solution to the problem of establishing interrater reliability.

Building on the work of Uebersax (2002) and J. R. Hayes and Hatch (1999), Stemler (2004) has argued that the wide variety of statistical techniques used for computing interrater reliability coefficients may be theoretically classified into one of three broad categories: (a) consensus estimates, (b) consistency estimates, and (c) measurement estimates. Statistics associated with these three categories differ in their assumptions about the purpose of the interrater reliability study, the nature of the data, and the implications for summarizing scores from various raters.

Consensus Estimates of Interrater Reliability

Consensus estimates are often used when one is attempting to demonstrate that a construct that traditionally has been considered highly subjective (e.g., creativity, wisdom, hate) can be reliably captured by independent raters. The assumption is that if independent raters are able to come to exact agreement about how to apply the various levels of a scoring rubric (which operationally defines behaviors associated with the construct), then this provides some defensible evidence for the existence of the construct. Furthermore, if two independent judges demonstrate high levels of agreement in their application of a scoring rubric to rate behaviors, then the two judges may be said to share a common interpretation of the construct.

Consensus estimates tend to be the most useful when data are nominal in nature and different levels of the rating scale represent qualitatively different ideas. Consensus estimates also can be useful when different levels of the rating scale are assumed to represent a linear continuum of the construct but are ordinal in nature (e.g., a Likert-type scale). In such cases, the judges must come to exact agreement about each of the quantitative levels of the construct under investigation.

The three most popular types of consensus estimates of interrater reliability found in the literature include (a) percent agreement and its variants, (b) Cohen's kappa and its variants (Agresti, 1996; Cohen, 1960, 1968; Krippendorff, 2004), and (c) odds ratios. Other less frequently used statistics that fall under this category include Jaccard's J and the G-Index (see Barrett, 2001).

Percent Agreement. Perhaps the most popular method for computing a consensus estimate of interrater reliability is through the use of the simple percent agreement statistic. For example, in a study examining creativity, Sternberg and Lubart (1995) asked sets of judges to rate the level of creativity associated with each of a number of products generated by study participants (e.g., draw a picture illustrating Earth from an insect's point of view, write an essay based on the title "2983"). The goal of their study was to demonstrate that creativity could be detected and objectively scored with high levels of agreement across independent judges. The authors reported percent agreement levels across raters of .92 (Sternberg & Lubart, 1995, p. 31).

The percent agreement statistic has several advantages. For example, it has a strong intuitive appeal, it is easy to calculate, and it is easy to explain. The statistic also has some distinct disadvantages, however. If the behavior of interest has a low or high incidence of occurrence in the population, then it is possible to get artificially inflated percent agreement figures simply because most of the values fall under one category of the rating scale (J. R. Hayes & Hatch, 1999). Another disadvantage to using the simple percent agreement figure is that it is often time-consuming and labor-intensive to train judges to the point of exact agreement.

One popular modification of the percent agreement figure found in the testing literature involves broadening the definition of agreement by including the adjacent scoring categories on the rating scale. For example, some testing programs include writing sections that are scored by judges using a rating scale with levels ranging from 1 (*low*) to 6 (*high*) (College Board, 2006). If a percent adjacent agreement approach were used to score this section of the exam, this would

mean that the judges would not need to come to exact agreement about the ratings they assign to each participant; rather, so long as the ratings did not differ by more than one point above or below the other judge, then the two judges would be said to have reached consensus. Thus, if Rater A assigns an essay a score of 3 and Rater B assigns the same essay a score of 4, the two raters are close enough together to say that they "agree," even though their agreement is not exact.

The rationale for the adjacent percent agreement approach is often a pragmatic one. It is extremely difficult to train independent raters to come to exact agreement, no matter how good one's scoring rubric. Yet, raters often give scores that are "pretty close" to the same, and we do not want to discard this information. Thus, the thinking is that if we have a situation in which two raters never differ by more than one score point in assigning their ratings, then we have a justification for taking the average score across all ratings. This logic holds under two conditions. First, the difference between raters must be randomly distributed across items. In other words, Rater A should not give systematically lower scores than Rater B. Second, the scores assigned by raters must be evenly distributed across all possible score categories. In other words, both raters should give equal numbers of 1s, 2s, 3s, 4s, 5s, and 6s across the population of essays that they have read. If both of these assumptions are met, then the adjacent percent agreement approach is defensible. If, however, either of these assumptions is violated, this could lead to a situation in which the validity of the resultant summary scores is dubious (see the box below).

Consider a situation in which Rater A systematically assigns scores that are one power lower than Rater B. Assume that they have each rated a common set of 100 essays. If we average the scores of the two raters across all essays to arrive at individual student scores, this seems, on the surface, to be defensible because it really does not matter whether Rater A or Rater B is assigning the high or low score because even if Rater A and Rater B had no systematic difference in severity of ratings, the average score would be the same. However, suppose that dozens of raters are used to score the essays. Imagine that Rater C is also called in to rate the same essay for a different sample of students. Rater C is paired up with Rater B within the context of an overlapping design to maximize rater efficiency (e.g., McArdle, 1994). Suppose that we find a situation in which Rater B is systematically lower than Rater C in assigning grades. In other words, Rater A is systematically one point lower than Rater B, and Rater B is systematically one point lower than Rater C.

On the surface, again, it seems logical to average the scores assigned by Rater B and Rater C. Yet, we now find ourselves in a situation in which the students rated by the Rater B/C pair score systematically one point higher than the students rated by the Rater A/B pair, even though neither combination of raters differed by more than one score point in their ratings, thereby demonstrating "interrater reliability." Which student would you rather be? The one who was lucky enough to draw the B/C rater combination or the one who unfortunately was scored by the A/B combination?

Thus, in order to make a validity argument for summarizing the results of multiple raters, it is not enough to demonstrate adjacent percent agreement between rater pairs; it must also be demonstrated that there is no systematic difference in rater severity between the rater set pairs.

This can be demonstrated (and corrected for in the final score) through the use of the many-facet Rasch model.

Now let us examine what happens if the second assumption of the adjacent percent agreement approach is violated. If you are a rater for a large testing company, and you are told that you will be retained only if you are able to demonstrate interrater reliability with everyone else, you would naturally look for your best strategy to maximize interrater reliability. If you are then told that your scores can differ by no more than one point from the other raters, you would quickly discover that your best bet then is to avoid giving any ratings at the extreme ends of the scale (i.e., a rating of 1 or a rating of 6). Why? Because a rating at the extreme end of the scale (e.g., 6) has two potential scores with which it can overlap (i.e., 5 or 6), whereas a rating of 5 would allow you to potentially "agree" with three scores

(Continued)

(Continued)

(i.e., 4, 5, or 6), thereby maximizing your chances of agreeing with the second rater. Thus, it is entirely likely that the scale will go from being a 6-point scale to a 4-point scale, reducing the overall variability in scores given across the spectrum of participants. If only four categories are used, then the percent agreement statistics will be artificially inflated due to chance factors. For example, when a scale is 1 to 6, two participants are expected to agree on ratings by chance alone only 17% of the time. When the scale is reduced to 1 to 4, the percent agreement expected by chance jumps to 25%. If three categories, a 33% chance agreement is expected; if two categories, a 50% chance agreement is expected. In other words, a 6-point scale that uses adjacent percent agreement scoring is most likely functionally equivalent to a 4-point scale that uses exact agreement scoring.

This approach is advantageous in that it relaxes the strict criterion that the judges agree exactly. On the other hand, percent agreement using adjacent categories can lead to inflated estimates of interrater reliability if there are only a limited number of categories to choose from (e.g., a 1–4 scale). If the rating scale has a limited number of points, then nearly all points will be adjacent, and it would be surprising to find agreement lower than 90%.

Cohen's Kappa. Another popular consensus estimate of interrater reliability is Cohen's kappa statistic (Cohen, 1960, 1968). Cohen's kappa was designed to estimate the degree of consensus between two judges and determine whether the level of agreement is greater than would be expected to be observed by chance alone (see Stemler, 2001, for a practical example with calculation). The interpretation of the kappa statistic is slightly different from the interpretation of the percent agreement figure (Agresti, 1996). A value of zero on kappa does not indicate that the two judges did not agree at all; rather, it indicates that the two judges did not agree with each other any more than would be predicted by chance alone. Consequently, it is possible to have negative values of kappa if judges agree less often than chance would predict. Kappa is a highly useful statistic when one is concerned that the percent agreement statistic may be artificially inflated due to the fact that most observations fall into a single category.

Kappa is often useful within the context of exploratory research. For example, Stemler and Bebell (1999) conducted a study aimed at detecting the various purposes of schooling articulated in school mission statements. Judges were given a scoring rubric that listed 10 possible thematic categories under which the main idea of each mission statement could be classified (e.g., social development, cognitive development, civic development). Judges then read a series of mission statements and attempted to classify each sampling unit according to the major purpose of schooling articulated. If both judges consistently rated the dominant theme of the mission statement as representing elements of citizenship, then they were said to have communicated with each other in a meaningful way because they had both classified the statement in the same way. If one judge classified the major theme as social development, and the other judge classified the major theme as citizenship, then a breakdown in shared understanding occurred. In that case, the judges were not coming to a consensus on how to apply the levels of the scoring rubric. The authors chose to use the kappa statistic to evaluate the degree of consensus because they did not expect the frequency of the major themes of the mission statements to be evenly distributed across the 10 categories of their scoring rubric.

Although some authors (Landis & Koch, 1977) have offered guidelines for interpreting kappa values, other authors (Krippendorff, 2004; Uebersax, 2002) have argued that the kappa values for different items or from different studies cannot be meaningfully compared unless the base rates are identical. Consequently, these authors suggest that although the statistic gives some indication as to whether the agreement is better than that predicted by chance alone, it is difficult to apply rules of thumb for interpreting kappa across different circumstances. Instead, Uebersax (2002) suggests that researchers using the kappa coefficient look at it

for up or down evaluation of whether ratings are different from chance, but they should not get too invested in its interpretation.

Krippendorff (2004) has introduced a new coefficient alpha into the literature that claims to be superior to kappa because alpha is capable of incorporating the information from multiple raters, dealing with missing data, and yielding a chance-corrected estimate of interrater reliability. The major disadvantage of Krippendorff's alpha is that it is computationally complex; however, statistical macros that compute Krippendorff's alpha have been created and are freely available (K. Hayes, 2006). In addition, however, some research suggests that in practice, alpha values tend to be nearly identical to kappa values (Dooley, 2006).

Odds Ratios. A third consensus estimate of interrater reliability is the odds ratio. The odds ratio is most often used in circumstances where raters are making dichotomous ratings (e.g., presence/absence of a phenomenon), although it can be extended to ordered category ratings. In a 2×2 contingency table, the odds ratio indicates how much the odds of one rater making a given rating (e.g., positive/negative) increase for cases when the other rater has made the same rating. For example, suppose that in a music competition with 100 contestants, Rater 1 gives 90 of them a positive score for vocal ability, while in the same sample of 100 contestants, Rater 2 only gives 20 of them a positive score for vocal ability. The odds of Rater 1 giving a positive vocal ability score are 90 to 10, or 9:1, while the odds of Rater 2 giving a positive vocal ability score are only 20 to 80, or 1:4 = 0.25:1. Now, 9/0.25 = 36, so the odds ratio is 36. Within the context of interrater reliability, the important idea captured by the odds ratio is whether it deviates substantially from 1.0. From the perspective of interrater reliability, it would be most desirable to have an odds ratio that is close to 1.0, which would indicate that Rater 1 and Rater 2 rated the same proportion of contestants as having high vocal ability. The larger the odds ratio value, the larger the discrepancy there is between raters in terms of their level of consensus.

The odds ratio has the advantage of being easy to compute and is familiar from other statistical applications (e.g., logistic regression). The disadvantage to the odds ratio is that it is most intuitive within the context of a 2×2 contingency table with dichotomous rating categories. Although the technique can be generalized to ordered category ratings, it involves extra computational complexity that undermines its intuitive advantage. Furthermore, as Osborne (2006) has pointed out, although the odds ratio is straightforward to compute, the interpretation of the statistic is not always easy to convey, particularly to a lay audience.

Computing Common Consensus Estimates of Interrater Reliability

Let us now turn to a practical example of how to calculate each of these coefficients. As an example data set, we will draw from Stemler, Grigorenko, Jarvin, and Sternberg's (2006) study in which they developed augmented versions of the Advanced Placement Psychology Examination. Participants were required to complete a number of essay items that were subsequently scored by different sets of raters. Essay Question 1, Part d was a question that asked participants to give advice to a friend who is having trouble sleeping, based on what they know about various theories of sleep. The item was scored using a 5-point scoring rubric. For this particular item, 75 participants received scores from two independent raters.

Percent Agreement. Percent agreement is calculated by adding up the number of cases that received the same rating by both judges and dividing that number by the total number of cases rated by the two judges. Using SPSS, one can run the crosstabs procedure and generate a table to facilitate the calculation (see Table 3.1). The percent agreement on this item is 42%; however, the percent adjacent agreement is 87%.

Cohen's Kappa. The formula for computing Cohen's kappa is listed in Formula 1.

$$\kappa = \frac{P_A - P_C}{1 - P_C}, \qquad (1)$$

where P_A = proportion of units on which the raters agree, and P_C = the proportion of units for which agreement is expected by chance.

It is possible to compute Cohen's kappa in SPSS by simply specifying in the crosstabs procedure the desire to produce Cohen's kappa (see Table 3.1). For this data set, the kappa value

Table 3.1 SPSS Code and Output for Percent Agreement and Percent Adjacent Agreement and Cohen's Kappa

SPSS CODE

```
CROSSTABS
/TABLES = Rater_1 BY Rater_2
/FORMAT = AVALUE TABLES
/STATISTIC = KAPPA
/CELLS = COUNT
/COUNT ROUND CELL
```

SPSS OUTPUT

		Rater_2					
		0	*1*	*2*	*3*	*4*	*Total*
Rater_1	0	1	0	0	0	0	1
	1	1	4	0	0	0	5
	2	1	8	3	1	0	13
	3	0	3	12	10	1	26
	4	0	0	6	11	13	30
Total		3	15	21	22	14	75

RESULTS

Percent agreement = 31/75 = 42%

Percent adjacent agreement = 65/75 = 87%

Cohen's kappa = .23

is .23, which indicates that the two raters agreed on the scoring only slightly more often than we would predict based on chance alone.

Odds Ratios. The formula for computing an odds ratio is shown in Formula 2.

	Rater_2		
Rater_1	+	−	
+	a	b	a + b
−	c	d	c + d
	a + c	b + d	Total

$$OR = \frac{[a/(a+b)]/[b/(a+b)]}{[c/(c+d)]/[d/(c+d)]} = \frac{a/b}{c/d} = \frac{ad}{bc}. \quad (2)$$

The SPSS code for computing the odds ratio is shown in Table 3.2. In order to compute the odds ratio using the crosstabs procedure in SPSS, it was necessary to recode the data so that the ratings were dichotomous. Consequently, ratings of 0, 1, and 2 were assigned a value of 0 (failing) while ratings of 3 and 4 were assigned a value of 1 (passing). The odds ratio for the current data set is 30, indicating that there was a substantial difference between the raters in terms of the proportion of students classified as passing versus failing.

Implications for Summarizing Scores From Various Raters

If raters can be trained to the point where they agree on how to assign scores from a rubric, then scores given by the two raters may be treated as equivalent. This fact has practical implications for determining the number of raters needed to complete a study. Thus, the remaining work of rating subsequent items can be split between the raters without both raters having to score all items. Furthermore, the

Table 3.2 SPSS Code and Output for Odds Ratios

SPSS CODE

```
CROSSTABS
/TABLES = Rater_1 BY Rater_2
/FORMAT = AVALUE TABLES
/STATISTIC = RISK
/CELLS = COUNT
/COUNT ROUND CELL
```

SPSS OUTPUT

		Rater_2		
		.00	*1.00*	*Total*
Rater_1	.00	18	1	19
	1.00	21	35	56
Total		39	36	75

		95% Confidence Interval	
	Value	*Lower*	*Upper*
Odds Ratio for Rater_1r (.00 / 1.00)	30.000	3.729	241.357
For cohort Rater_2r = .00	2.526	1.773	3.601
For cohort Rater_2r = 1.00	.084	.012	.574
N of Valid Cases	75		

RESULTS

Odds ratio for Rater 1 (0/1) = 30

summary scores may be calculated by simply taking the score from one of the judges or by averaging the scores given by all of the judges, since high interrater reliability indicates that the judges agree about how to apply the rating scale. A typical guideline found in the literature for evaluating the quality of interrater reliability based on consensus estimates is that they should be 70% or greater. If raters are shown to reach high levels of consensus, then adding more raters adds little extra information from a statistical perspective and is probably not justified from the perspective of resources.

Advantages of Consensus Estimates

One particular advantage of the consensus approach to estimating interrater reliability is that the calculations are easily done by hand. A second advantage is that the techniques falling within this general category are well suited to dealing with nominal variables whose levels on the rating scale represent qualitatively different categories. A third advantage is that consensus estimates can be useful in diagnosing problems with judges' interpretations of how to apply the rating scale. For example, inspection of the information from a crosstab table may allow the researcher to realize that the judges may be unclear about the rules for when they are supposed to score an item as zero as opposed to when they are supposed to score the item as missing. A visual analysis of the output allows the researcher to go back to the data and clarify the discrepancy or retrain the judges.

When judges exhibit a high level of consensus, it implies that both judges are essentially providing the same information. One implication of a high

consensus estimate of interrater reliability is that both judges need not score all remaining items. For example, if there were 100 tests to be scored after the interrater reliability study was finished, it would be most efficient to ask Judge A to rate exams 1 to 50 and Judge B to rate exams 51 to 100 because the two judges have empirically demonstrated that they share a similar meaning for the scoring rubric. In practice, however, it is usually a good idea to build in a 30% overlap between judges even after they have been trained, in order to provide evidence that the judges are not drifting from their consensus as they read more items.

Disadvantages of Consensus Estimates

One disadvantage of consensus estimates is that interrater reliability statistics must be computed separately for each item and for each pair of judges. Consequently, when reporting consensus-based interrater reliability estimates, one should report the minimum, maximum, and median estimates for all items and for all pairs of judges.

A second disadvantage is that the amount of time and energy it takes to train judges to come to exact agreement is often substantial, particularly in applications where exact agreement is unnecessary (e.g., if the exact application of the levels of the scoring rubric is not important, but rather a means to the end of getting a summary score for each respondent).

Third, as Linacre (2002) has noted, training judges to a point of forced consensus may actually reduce the statistical independence of the ratings and threaten the validity of the resulting scores.

Finally, consensus estimates can be overly conservative if two judges exhibit systematic differences in the way that they use the scoring rubric but simply cannot be trained to come to a consensus. As we will see in the next section, it is possible to have a low consensus estimate of interrater reliability while having a high consistency estimate and vice versa. Consequently, sole reliance on consensus estimates of interrater reliability might lead researchers to conclude that "interrater reliability is low" when it may be more precisely stated that the *consensus estimate* of interrater reliability is low.

Consistency Estimates of Interrater Reliability

Consistency estimates of interrater reliability are based on the assumption that it is not really necessary for raters to share a common interpretation of the rating scale, so long as each judge is consistent in classifying the phenomenon according to his or her own definition of the scale. For example, if Rater A assigns a score of 3 to a certain group of essays, and Rater B assigns a score of 1 to that same group of essays, the two raters have not come to a consensus about how to apply the rating scale categories, but the difference in how they apply the rating scale categories is predictable.

Consistency approaches to estimating interrater reliability are most useful when the data are continuous in nature, although the technique can be applied to categorical data if the rating scale categories are thought to represent an underlying continuum along a unidimensional construct. Values greater than .70 are typically acceptable for consistency estimates of interrater reliability (Barrett, 2001).

The three most popular types of consistency estimates are (a) correlation coefficients (e.g., Pearson, Spearman), (b) Cronbach's alpha (Cronbach, 1951), and (c) intraclass correlation. For information regarding additional consistency estimates of interrater reliability, see Bock, Brennan, and Muraki (2002); Burke and Dunlap (2002); LeBreton, Burgess, Kaiser, Atchley, and James (2003); and Uebersax (2002).

Correlation Coefficients. Perhaps the most popular statistic for calculating the degree of consistency between raters is the Pearson correlation coefficient. Correlation coefficients measure the association between independent raters. Values approaching +1 or −1 indicate that the two raters are following a systematic pattern in their ratings, while values approaching zero indicate that it is nearly impossible to predict the score one rater would give by knowing the score the other rater gave. It is important to note that even though the correlation between scores assigned by two judges may be nearly perfect, there may be substantial mean differences between the raters. In other words, two raters may differ in the absolute values they assign to each rating by two points; however, so long as there is a 2-point difference for each rating they assign, the raters will have achieved high consistency estimates of interrater reliability. Thus, a large value for a measure of association does not imply that the raters are agreeing on the actual application of the rating scale, only that they are consistent in applying the ratings according to their own unique understanding of the scoring rubric.

The Pearson correlation coefficient can be computed by hand (Glass & Hopkins, 1996) or can easily be computed using most statistical packages. One beneficial feature of the Pearson correlation coefficient is that the scores on the rating scale can be continuous in nature (e.g., they can take on partial values such as 1.5). Like the percent agreement statistic, the Pearson correlation coefficients can be calculated only for one pair of judges at a time and for one item at a time.

A potential limitation of the Pearson correlation coefficient is that it assumes that the data underlying the rating scale are normally distributed. Consequently, if the data from the rating scale tend to be skewed toward one end of the distribution, this will attenuate the upper limit of the correlation coefficient that can be observed. The Spearman rank coefficient provides an approximation of the Pearson correlation coefficient but may be used in circumstances where the data under investigation are not normally distributed. For example, rather than using a continuous rating scale, each judge may rank order the essays that he or she has scored from best to worst. In this case, then, since both ratings being correlated are in the form of rankings, a correlation coefficient can be computed that is governed by the number of pairs of ratings (Glass & Hopkins, 1996). The major disadvantage to Spearman's rank coefficient is that it requires both judges to rate all cases.

Cronbach's Alpha. In situations where more than two raters are used, another approach to computing a consistency estimate of interrater reliability would be to compute Cronbach's alpha coefficient (Crocker & Algina, 1986). Cronbach's alpha coefficient is a measure of internal consistency reliability and is useful for understanding the extent to which the ratings from a group of judges hold together to measure a common dimension. If the Cronbach's alpha estimate among the judges is low, then this implies that the majority of the variance in the total composite score is really due to error variance and not true score variance (Crocker & Algina, 1986).

The major advantage of using Cronbach's alpha comes from its capacity to yield a single consistency estimate of interrater reliability across multiple judges. The major disadvantage of the method is that each judge must give a rating on every case, or else the alpha will only be computed on a subset of the data. In other words, if just one rater fails to score a particular individual, that individual will be left out of the analysis. In addition, as Barrett (2001) has noted, "because of this 'averaging' of ratings, we reduce the variability of the judges' ratings such that when we average all judges' ratings, we effectively remove all the error variance for judges" (p. 7).

Intraclass Correlation. A third popular approach to estimating interrater reliability is through the use of the intraclass correlation coefficient. An interesting feature of the intraclass correlation coefficient is that it confounds two ways in which raters differ: (a) consensus (or bias—i.e., mean differences) and (b) consistency (or association). As a result, the value of the intraclass correlation coefficient will be decreased in situations where there is a low correlation between raters *and* in situations where there are large mean differences between raters. For this reason, the intraclass correlation may be considered a conservative estimate of interrater reliability. If the intraclass correlation coefficient is close to 1, then chances are good that this implies that excellent interrater reliability has been achieved.

The major advantage of the intraclass correlation is its capacity to incorporate information from different types of rater reliability data. On the other hand, as Uebersax (2002) has noted, "If the goal is to give feedback to raters to improve future ratings, one should distinguish between these two sources of disagreement" (p. 5). In addition, because the intraclass correlation represents the ratio of within-subject variance to between-subject variance on a rating scale, the results may not look the same if raters are rating a homogeneous subpopulation as opposed to the general population. Simply by restricting the between-subject variance, the intraclass correlation will be lowered. Therefore, it is important to pay special attention to the population being assessed and to understand that this can influence the value of the intraclass correlation coefficient (ICC). For this reason, ICCs are not directly comparable across populations. Finally, it is important to note that, like the Pearson correlation coefficient, the intraclass correlation coefficient will be attenuated if assumptions of normality in rating data are violated.

Computing Common Consistency Estimates of Interrater Reliability

Let us now turn to a practical example of how to calculate each of these coefficients. We will use the same data set and compute each estimate on the data.

Correlation Coefficients. The formula for computing the Pearson correlation coefficient is listed in Formula 3.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}. \quad (3)$$

Using SPSS, one can run the correlate procedure and generate a table similar to Table 3.3. One may request both Pearson and Spearman correlation coefficients. The Pearson correlation coefficient on this data set is .76; the Spearman correlation coefficient is .74.

Cronbach's Alpha. The Cronbach's alpha value is calculated using Formula 4,

$$\alpha = \frac{N}{N-1}\left(\frac{\sigma_x^2 - \sum_{i=1}^{N} \sigma^2 Y_i}{\sigma_x^2}\right), \quad (4)$$

where

N is the number of components (raters),

σ_x^2 is the variance of the observed total scores, and

$\sigma^2 Y_i$ is the variance of component i.

In order to compute Cronbach's alpha using SPSS, one may simply specify in the crosstabs procedure the desire to produce Cronbach's alpha (see Table 3.4). For this example, the alpha value is .86.

Table 3.3 SPSS Code and Output for Pearson and Spearman Correlations

SPSS CODE

CROSSTABS

/TABLES = Rater_1 BY Rater_2

/FORMAT = AVALUE TABLES

/STATISTIC = CORR KAPPA

/CELLS = COUNT

/COUNT ROUND CELL

SPSS OUTPUT

		Value	*Asymp. Std. Error*[a]	*Approx. T*[b]	*Asymp. Sig.*
Interval by Interval	Pearson's *R*	.761	.044	10.027	.000[c]
Ordinal by Ordinal	Spearman Correlation	.744	.057	9.504	.000[c]
Measure of Agreement	Kappa	.229	.071	3.765	.000
N of Valid Cases		75			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

RESULTS

Pearson correlation r = .76

Spearman correlation rho = .74

Table 3.4 SPSS Code and Output for Cronbach's Alpha

SPSS CODE

RELIABILITY

/VARIABLES = Rater_1 Rater_2

/SCALE('ALL VARIABLES') ALL/MODEL = ALPHA

SPSS OUTPUT

Cronbach's Alpha	N *of Items*
.860	2

RESULTS

Cronbach's alpha = .86

Intraclass Correlation. Formula 5 presents the equation used to compute the intraclass correlation value.

$$ICC = \frac{\sigma^2(b)}{\sigma^2(b) + \sigma^2(w)}, \qquad (5)$$

where

$\sigma^2(b)$ is the variance of the ratings between judges, and

$\sigma^2(w)$ is the pooled variance within raters.

In order to compute intraclass correlation, one may specify the procedure in SPSS using the code listed in Table 3.5. The intraclass correlation coefficient for this data set is .75.

Implications for Summarizing Scores From Various Raters

It is important to recognize that although consistency estimates may be high, the means and medians of the different judges may be very different. Thus, if one judge consistently gives scores that are 2 points lower on the rating scale than does a second judge, the scores will ultimately need to be corrected for this difference in judge severity if the final scores are to be summarized or subjected to further analyses.

Table 3.5 SPSS Code and Output for Intraclass Correlation

SPSS CODE

RELIABILITY

/VARIABLES = Rater_1 Rater_2

/SCALE('ALL VARIABLES') ALL/MODEL = ALPHA

/ICC = MODEL(MIXED) TYPE(CONSISTENCY) CIN = 95 TESTVAL = 0

SPSS OUTPUT

	Intraclass Correlation[a]	*95% Confidence Interval*		*F Test With True Value 0*			
		Lower Bound	*Upper Bound*	*Value*	df1	df2	*Sig.*
Single Measures	.754[b]	.637	.837	7.143	74.0	74	.000
Average Measures	.860[c]	.778	.912	7.143	74.0	74	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition; the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent because it is not estimable otherwise.

RESULTS

Intraclass correlation = .75

Advantages of Consistency Estimates

There are three major advantages to using consistency estimates of interrater reliability. First, the approach places less stringent demands on the judges in that they need not be trained to come to exact agreement with one another so long as each judge is consistent within his or her own definition of the rating scale (i.e., exhibits high intrarater reliability). It is sometimes the case that the exact application of the levels of the scoring rubric is not important in itself. Instead, the scoring rubric is a means to the end of creating scores for each participant that can be summarized in a meaningful way. If summarization is the goal, then what is most important is that each judge apply the rating scale consistently within his or her own definition of the rating scale, regardless of whether the two judges exhibit exact agreement. Consistency estimates allow for the detection of systematic differences between judges, which may then be adjusted statistically. For example, if Judge A consistently gives scores that are 2 points lower than Judge B does, then adding 2 extra points to the exams of all students who were scored by Judge A would provide an equitable adjustment to the raw scores.

A second advantage of consistency estimates is that certain methods within this category (e.g., Cronbach's alpha) allow for an overall estimate of consistency among multiple judges. The third advantage is that consistency estimates readily handle continuous data.

Disadvantage of Consistency Estimates

One disadvantage of consistency estimates is that if the construct under investigation has some objective meaning, then it may not be desirable for the two judges to "agree to disagree." Instead, it may be important for the judges to come to an exact agreement on the scores that they are generating.

A second disadvantage of consistency estimates is that judges may differ not only systematically in the raw scores they apply but also in the number of rating scale categories they use. In that case, a mean adjustment for a severe judge may provide a partial solution, but the two judges may also differ on the variability in scores they give. Thus, a mean adjustment alone will not effectively correct for this difference.

A third disadvantage of consistency estimates is that they are highly sensitive to the distribution of the observed data. In other words, if most of the ratings fall into one or two categories, the correlation coefficient will necessarily be deflated due to restricted variability. Consequently, a reliance on the consistency estimate alone may lead the researcher to falsely conclude that interrater reliability was poor without specifying more precisely that the *consistency estimate* of interrater reliability was poor and providing an appropriate rationale.

Measurement Estimates of Interrater Reliability

Measurement estimates are based on the assumption that one should use all of the information available from all judges (including discrepant ratings) when attempting to create a summary score for each respondent. In other words, each judge is seen as providing some unique information that is useful in generating a summary score for a person. As Linacre (2002) has noted, "It is the accumulation of information, not the ratings themselves, that is decisive" (p. 858). Consequently, under the measurement approach, it is not necessary for two judges to come to a consensus on how to apply a scoring rubric because differences in judge severity can be estimated and accounted for in the creation of each participant's final score.

Measurement estimates are also useful in circumstances where multiple judges are providing ratings, and it is impossible for all judges to rate all items. They are best used when different levels of the rating scale are intended to represent different levels of an underlying unidimensional construct (e.g., mathematical competence).

The two most popular types of measurement estimates are (a) factor analysis and (b) the many-facets Rasch model (Linacre, 1994; Linacre, Englehard, Tatem, & Myford, 1994; Myford & Cline, 2002) or log-linear models (von Eye & Mun, 2004).

Factor Analysis. One popular measurement estimate of interrater reliability is computed using factor analysis (Harman, 1967). Using this method, multiple judges may rate a set of participants. The judges' scores are then subjected to a common factor analysis in order to determine the amount of shared variance in the ratings

that could be accounted for by a single factor. The percentage of variance that is explainable by the first factor gives some indication of the extent to which the multiple judges are reaching agreement. If the shared variance is high (e.g., greater than 60%), then this gives some indication that the judges are rating a common construct. The technique can also be used to check the extent to which judges agree on the number of underlying dimensions in the data set.

Once interrater reliability has been established in this way, each participant may then receive a single summary score corresponding to his or her loading on the first principal component underlying the set of ratings. This score can be computed automatically by most statistical packages.

The advantage of this approach is that it assigns a summary score for each participant that is based only on the relevance of the strongest dimension underlying the data. The disadvantage to the approach is that it assumes that ratings are assigned without error by the judges.

Many-Facets Rasch Measurement and Log-Linear Models. A second measurement approach to estimating interrater reliability is through the use of the many-facets Rasch model (Linacre, 1994).[2] Recent advances in the field of measurement have led to an extension of the standard Rasch measurement model (Rasch, 1960/1980; Wright & Stone, 1979). This new, extended model, known as the many-facets Rasch model, allows judge severity to be derived using the same scale (i.e., the logit scale) as person ability and item difficulty. In other words, rather than simply assuming that a score of 3 from Judge A is equally difficult for a participant to achieve as a score of 3 from Judge B, the equivalence of the ratings between judges can be empirically determined. Thus, it could be the case that a score of 3 from Judge A is really closer to a score of 5 from Judge B (i.e., Judge A is a more severe rater). Using a many-facets analysis, each essay item or behavior that was rated can be directly compared.

In addition, the difficulty of each item, as well as the severity of all judges who rated the items, can also be directly compared. For example, if a history exam included five essay questions and each of the essay questions was rated by 3 judges (2 unique judges per item and 1 judge who scored all items), the facets approach would allow the researcher to directly compare the severity of a judge who rated only Item 1 with the severity of a judge who rated only Item 4. Each of the 11 judges (2 unique judges per item + 1 judge who rated all items = 5*2 + 1 = 11) could be directly compared. The mathematical representation of the many-facets Rasch model is fully described in Linacre (1994).

Finally, in addition to providing information that allows for the evaluation of the severity of each judge in relation to all other judges, the facets approach also allows one to evaluate the extent to which each of the individual judges is using the scoring rubric in a manner that is internally consistent (i.e., an estimate of intrarater reliability). In other words, even if judges differ in their interpretation of the rating scale, the fit statistics will indicate the extent to which a given judge is faithful to his or her own definition of the scale categories across items and people.

The many-facets Rasch approach has several advantages. First, the technique puts rater severity on the same scale as item difficulty and person ability (i.e., the logit scale). Consequently, this feature allows for the computation of a single final summary score that is already corrected for rater severity. As Linacre (1994) has noted, this provides a distinct advantage over generalizability studies since the goal of a generalizability study is to determine

> the error variance associated with each judge's ratings, so that correction can be made to ratings awarded by a judge when he is the only one to rate an examinee. For this to be useful, examinees must be regarded as randomly sampled from some population of examinees which means that there is no way to correct an individual examinee's score for judge behavior, in a way which would be helpful to an examining board. This approach, however, was developed for use in contexts in which only estimates of population parameters are of interest to researchers. (p. 29)

Second, the item fit statistics provide some estimate of the degree to which each individual rater was applying the scoring rubric in an internally consistent manner. In other words, high-fit statistic values are an indication of rater drift over time.

Third, the technique works with multiple raters and does not require all raters to evaluate all objects. In other words, the technique is well suited to overlapping research designs, which

allows the researcher to use resources more efficiently. So long as there is sufficient connectedness in the data set (Engelhard, 1997), the severity of all raters can be evaluated relative to each other.

The major disadvantage to the many-facets Rasch approach is that it is computationally intensive and therefore is best implemented using specialized statistical software (Linacre, 1988). In addition, this technique is best suited to data that are ordinal in nature.

Computing Common Measurement Estimates of Interrater Reliability

Measurement estimates of interrater reliability tend to be much more computationally complex than consensus or consistency estimates. Consequently, rather than present the detailed formulas for each technique in this section, we instead refer to some excellent sources that are devoted to fully expounding the detailed computations involved. This will allow us to focus on the interpretation of the results of each of these techniques.

Factor Analysis. The mathematical formulas for computing factor-analytic solutions are expounded in several excellent texts (e.g., Harman, 1967; Kline, 1998). When using factor analysis to estimate interrater reliability, the data set should be structured in such a way that each column in the data set corresponds to the score given by Rater X on Item Y to each object in the data set (objects each receive their own row). Thus, if five raters were to score three essays from 100 students, the data set should contain 15 columns (e.g., Rater1_Item1, Rater2_Item1, Rater1_Item2) and 100 rows. In this example, we would run a separate factor analysis for each essay item (e.g., a 5×100 data matrix). Table 3.6 shows the SPSS code and output for running the factor analysis procedure.

There are two important pieces of information generated by the factor analysis. The first important piece of information is the value of the explained variance in the first factor. In the example output, the shared variance of the first factor is 76%, indicating that independent raters agree on the underlying nature of the construct being rated, which is also evidence of interrater reliability. In some cases, it may turn out that the variance in ratings is distributed over more than one factor. If that is the case, then this provides some evidence to suggest that the raters are not interpreting the underlying construct in the same manner (e.g., recall the example about creativity mentioned earlier in this chapter).

The second important piece of information comes from the factor loadings. Each object that has been rated will have a loading on each underlying factor. Assuming that the first factor explains most of the variance, the score to be used in subsequent analyses should be the loading on the primary factor.

Many-Facets Rasch Measurement. The mathematical formulas for computing results using the many-facets Rasch model may be found in Linacre (1994). In practice, the many-facets Rasch model is best implemented through the use of specialized software (Linacre, 1988). An example output of a many-facets Rasch analysis is listed in Table 3.7. The example output presented here is derived from the larger Stemler et al. (2006) data set.

The key values to interpret within the context of the many-facets Rasch approach are rater severity measures and fit statistics. Rater severity indices are useful for estimating the extent to which systematic differences exist between raters with regard to their level of severity. For example, rater CL was the most severe rater, with an estimated severity measure of +0.89 logits. Consequently, students whose test items were scored by CL would be more likely to receive lower raw scores than students who had the same test item scored by any of the other raters used in this project. At the other extreme, rater AP was the most lenient rater, with a rater severity measure of −0.91 logits. Consequently, simply using raw scores would lead to biased estimates of student proficiency since student estimates would depend, to an important degree, on which rater scored their essay. The facets program corrects for these differences and incorporates them into student ability estimates. If these differences were not taken into account when calculating student ability, students who had their exams scored by AP would be more likely to receive substantially higher raw scores than if the same item were rated by any of the other raters.

The results presented in Table 3.7 show that there is about a 1.5-logit spread in systematic

Table 3.6 SPSS Code and Output for Factor Analysis

SPSS CODE

```
FACTOR
/VARIABLES Rater_1 Rater_2 /MISSING LISTWISE /ANALYSIS Rater_1 Rater_2
/PRINT INITIAL EXTRACTION
/CRITERIA FACTORS(1) ITERATE(25)
/EXTRACTION PAF
/ROTATION NOROTATE
/SAVE REG(ALL)
/METHOD = CORRELATION
```

SPSS OUTPUT

	Initial	*Extraction*
Rater_1	.579	.760
Rater_2	.579	.760

Extraction Method: Principal Axis Factoring.

Total Variance Explained

	Initial Eigenvalues			*Extraction Sums of Squared Loadings*		
Factor	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	1.761	88.057	88.057	1.521	76.043	76.043
2	.239	11.943	100.000			

Extraction Method: Principal Axis Factoring.

RESULTS

Shared variance between raters = 76%

differences in rater severity (from −0.91 to +0.89). Consequently, assuming that all raters are defining the rating scales they are using in the same way is not a tenable assumption, and differences in rater severity must be taken into account in order to come up with precise estimates of student ability.

In addition to providing information that allows us to evaluate the severity of each rater in relation to all other raters, the facets approach also allows us to evaluate the extent to which each of the individual raters is using the scoring rubric in a manner that is internally consistent (i.e., intrarater reliability). In other words, even if raters differ in their own definition of how they use the scale, the fit statistics will indicate the extent to which a given rater is faithful to his or her own definition of the scale categories across items and people. Rater fit statistics are presented in columns 5 and 6 of Table 3.7.

Fit statistics provide an empirical estimate of the extent to which the expected response patterns for each individual match the observed response patterns. These fit statistics are interpreted much the same way as item or person infit statistics are interpreted (Bond & Fox, 2001; Wright & Stone, 1979). An infit value greater than 1.4 indicates that there is 40% more variation in the data than predicted by the Rasch model. Conversely, an infit value of 0.5 indicates that there is 50% less

Table 3.7 Output for a Many-Facets Rasch Analysis

Rater	*N of ratings*	*Measure*	*SE*	*InfitMS*	*OutfitMS*
CL	258	0.89	0.04	3.4	3.3
SK	1,896	0.38	0.01	1.4	1.6
AM	1,491	0.14	0.02	2.2	2.1
JJ	756	0.11	0.02	1.3	1.4
LM	758	0.09	0.02	1.2	1.3
CK	956	0.03	0.02	1.0	1.0
SG	815	0.02	0.02	1.9	2.0
KV	2,685	0.01	0.01	1.6	1.6
ER	3,369	−0.02	0.01	1.0	1.1
JW	940	−0.17	0.02	2.4	2.5
KA	1,662	−0.57	0.02	1.1	1.0
AP	869	−0.91	0.03	1.0	0.9

variation in the data than predicted by the Rasch model. Infit mean squares that are greater than 1.3 indicate that there is more unpredictable variation in the raters' responses than we would expect based on the model. Infit mean square values that are less than 0.7 indicate that there is less variation in the raters' responses than we would predict based on the model. Myford and Cline (2002) note that high infit values may suggest that ratings are noisy as a result of the raters' overuse of the extreme scale categories (i.e., the lowest and highest values on the rating scale), while low infit mean square indices may be a consequence of overuse of the middle scale categories (e.g., moderate response bias).

> The infit and outfit mean-square indices are unstandardized, information-weighted indices; by constrast the infit and outfit standardized indices are unweighted indices that are standardized toward a unit-normal distribution. These standardized indices are sensitive to sample size and, consequently, the accuracy of the standardization is data dependent. The expectation for the mean square index is 1.0; the range is 0 to infinity (Myford & Cline, 2002, p. 14).

The results in Table 3.7 reveal that 6 of the 12 raters had infit mean-square indices that exceeded 1.3. Raters CL (infit of 3.4), JW (infit of 2.4), and AM (infit of 2.2) appear particularly problematic. Their high infit values suggest that these raters are not using the scoring rubrics in a consistent way. The table of misfitting ratings provided by the facets computer program output allowed for an investigation of the exact nature of the highly unexpected response patterns associated with each of these raters. The table of misfitting ratings provides information on discrepant ratings based on two criteria: (a) how the other raters scored the item and (b) the particular raters' typical level of severity in scoring items of similar difficulty.

Implications for Summarizing Scores From Various Raters

Measurement estimates allow for the creation of a summary score for each participant that represents that participant's score on the underlying factor of interest, taking into account the extent to which each judge influences the score.

Advantages of Measurement Estimates

There are several advantages to estimating interrater reliability using the measurement approach. First, measurement estimates can take into account errors at the level of each judge or for groups of judges. Consequently, the summary scores generated from measurement

estimates of interrater reliability tend to more accurately represent the underlying construct of interest than do the simple raw score ratings from the judges.

Second, measurement estimates effectively handle ratings from multiple judges by simultaneously computing estimates across all of the items that were rated, as opposed to calculating estimates separately for each item and each pair of judges.

Third, measurement estimates have the distinct advantage of not requiring all judges to rate all items in order to arrive at an estimate of interrater reliability. Rather, judges may rate a particular subset of items, and as long as there is sufficient connectedness (Linacre, 1994; Linacre et al., 1994) across the judges and ratings, it will be possible to directly compare judges.

Disadvantages of Measurement Estimates

The major disadvantage of measurement estimates is that they are unwieldy to compute by hand. Unlike the percent agreement figure or correlation coefficient, measurement approaches typically require the use of specialized software to compute.

A second disadvantage is that certain methods for computing measurement estimates (e.g., facets) can handle only ordinal-level data. Furthermore, the file structure required to use facets is somewhat counterintuitive.

Summary and Conclusion

In this chapter, we have attempted to outline a framework for thinking about interrater reliability as a multifaceted concept. Consequently, we believe that there is no silver bullet "best" approach for its computation. There are multiple techniques for computing interrater reliability, each with its own assumptions and implications. As Snow, Cook, Lin, Morgan, and Magaziner (2005) have noted, "Percent/proportion agreement is affected by chance; kappa and weighted kappa are affected by low prevalence of condition of interest; and correlations are affected by low variability, distribution shape, and mean shifts" (p. 1682). Yet each technique (and class of techniques) has its own strengths and weaknesses.

Consensus estimates of interrater reliability (e.g., percent agreement, Cohen's kappa, odds ratios) are generally easy to compute and useful for diagnosing rater disparities; however, training raters to exact consensus requires substantial time and energy and may not be entirely necessary, depending on the goals of the study.

Consistency estimates of interrater reliability (e.g., Pearson and Spearman correlations, Cronbach's alpha, and intraclass correlations) are familiar and fairly easy to compute. They have the additional advantage of not requiring raters to perfectly agree with each other but only require consistent application of a scoring rubric within raters—systematic variance between raters is easily tolerated. The disadvantage to consistency estimates, however, is that they are sensitive to the distribution of the data (the more it departs from normality, the more attenuated the results). Furthermore, even if one achieves high consistency estimates, further adjustment to an individual's raw scores may be required in order to arrive at an unbiased final score that may be used in subsequent data analyses.

Measurement estimates of interrater reliability (e.g., factor analysis, many-facets Rasch measurement) can deal effectively with multiple raters, easily derive adjusted summary scores that are corrected for rater severity, and allow for highly efficient designs (e.g., not all raters need to rate all objects); however, this comes at the expense of added computational complexity and increased demands on resources (e.g., time and expertise).

In the end, the best technique will always depend on (a) the goals of the analysis (e.g., the stakes associated with the study outcomes), (b) the nature of the data, and (c) the desired level of information based on the resources available. The answers to these three questions will help to determine how many raters one needs, whether the raters need to be in perfect agreement with each other, and how to approach creating summary scores across raters.

We conclude this chapter with a brief table that is intended to provide rough interpretive guidance with regard to acceptable interrater reliability values (see Table 3.8). These values simply represent conventions the authors have encountered in the literature and via discussions with colleagues and reviewers; however, keep in mind that these guidelines are just rough estimates and will vary depending on the purpose of the study and the stakes associated with the

outcomes. The conventions articulated here assume that the interrater reliability study is part of a low-stakes, exploratory research study.

Table 3.8 General Guidelines for Interpreting Various Interrater Reliability Coefficients

Consensus Estimates	*Acceptable*
Percent agreement	70%
Cohen's kappa	0.50
Odds ratio	Close to 1.0
Consistency Estimates	*Acceptable*
Pearson correlation	0.70
Cronbach's alpha	0.70
Intraclass correlation	0.60
Measurement Estimates	*Acceptable*
Factor analysis	70% explained variance
Many-facets Rasch	.70 < rater infit values < 1.3

NOTE: The odds ratio should always be computed so that the outcome is greater than 1.0.

Notes

1. Also known as interobserver or interjudge reliability or agreement.
2. Readers interested in this model can refer to Chapters 4 and 5 on Rasch measurement for more information.

References

Agresti, A. (1996). *An introduction to categorical data analysis* (2nd ed.). New York: John Wiley.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Barrett, P. (2001, March). *Assessing the reliability of rating data.* Retrieved June 16, 2003, from http://www.liv.ac.uk/~pbarrett/rater.pdf

Bock, R., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*(4), 364–375.

Bond, T., & Fox, C. (2001). *Applying the Rasch model.* Mahwah, NJ: Lawrence Erlbaum.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*(2), 159–172.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scale disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

College Board. (2006). *How the essay is scored.* Retrieved November 4, 2006, from http://www.collegeboard.com/student/testing/sat/about/sat/essay_scoring.html

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Harcourt Brace Jovanovich.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Dooley, K. (2006). *Questionnaire Programming Language—Interrater reliability report.* Retrieved November 4, 2006, from http://qpl.gao.gov/ca050404.htm

Engelhard, G. (1997). Constructing rater and task banks for performance assessment. *Journal of Outcome Measurement, 1*(1), 19–33.

Glass, G. V., & Hopkins, K. H. (1996). *Statistical methods in education and psychology.* Boston: Allyn & Bacon.

Harman, H. H. (1967). *Modern factor analysis.* Chicago: University of Chicago Press.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*(3), 354–367.

Hayes, K. (2006). *SPSS Macro for computing Krippendorff's alpha.* Retrieved from http://www.comm.ohio-state.edu/ahayes/SPSS%20programs/kalpha.htm

Hopkins, K. H. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn & Bacon.

Kline, R. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

LeBreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability

and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*(1), 80–128.

Linacre, J. M. (1988). FACETS: A computer program for many-facet Rasch measurement (Version 3.3.0). Chicago: MESA Press.

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (2002). Judge ratings with forced agreement. *Rasch Measurement Transactions, 16*(1), 857–858.

Linacre, J. M., Englehard, G., Tatem, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*(4), 569–577.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29*(4), 409–454.

Myford, C. M., & Cline, F. (2002, April 1–5). *Looking for patterns in disagreements: A facets analysis of human raters' and e-raters' scores on essays written for the Graduate Management Admission Test (GMAT).* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Osborne, J. W. (2006). Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses. *Practical Assessment, Research & Evaluation, 11*(7). Retrieved from http://pareonline.net/getvn.asp?v=11&n=17

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press. (Original work published 1960)

Snow, A. L., Cook, K. F., Lin, P.-S., Morgan, R. O., & Magaziner, J. (2005). Proxies and other external raters: Methodological considerations. *Health Services Research, 40*(5), 1676–1693.

Stemler, S. E. (2001). An overview of content analysis. *Practical Assessment, Research and Evaluation, 7*(17). Retrieved from http://ericae.net/pare/getvn.asp?v=7&n=17

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Stemler, S. E., & Bebell, D. (1999, April). *An empirical approach to understanding and analyzing the mission statements of selected educational institutions.* Paper presented at the New England Educational Research Organization (NEERO), Portsmouth, NH.

Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2006). Using the theory of successful intelligence as a basis for augmenting AP exams in psychology and statistics. *Contemporary Educational Psychology, 31*(2), 75–108.

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity.* New York: Free Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Uebersax, J. (2002). *Statistical methods for rater agreement.* Retrieved August 9, 2002, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm

von Eye, A., & Mun, E. Y. (2004). *Analyzing rater agreement: Manifest variable methods.* Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.